Predictive Analysis

by

Seymour Geisser

University of Minnesota

Technical Report No. 415

January, 1983

Supported in part by NIH Grant GM25271.

One of the curious turns taken by theoreticians in statistics during this century was to change the orientation of the field from an analysis of functions of finite observables (observabilism, predictivism) to one on unknown and unknowable parameters. An unknowable parameter is one whose exact value can never be ascertained from a finite number of observations or measurements no matter how large. The analysis of observables such as estimating the total or average response of a finite population from a sample and similar problems had been the major focus of attention of statisticians up until this century - so that Karl Pearson (1907) could state that the fundamental problem of statistics was predictive. The introduction of mathematically defined models indexed by parameters and Fisher's clear distinction of parameters and statistics initiated the stress on parametric inference. Possibly because of the attractive mathematics, hypotheses testing and the estimation of parameters, though speedily outgrowing their potential applicability, completely absorbed the attention of mathematical statisticians until rather recently. It has become clear that for the most part parametric analysis as such can be viewed as a special or limiting case of the predictivistic or observabilistic approach, Geisser (1982). In most statistical applications there are two basic types of models - the error (or measurement) model e.g. $X = \theta + e$ where θ is a true value, say some physical entity like the weight of a rock, real and observable but imperfectly measured. Although θ here is generally viewed as a parameter it is in an extended sense an observable entity (not an index of a population except modeled as such for convenience). In this instance, it will not matter whether we ascribe this to observable or parametric inference. But this model occurs far less frequently than one wherein a sample of units drawn from some population is measured with respect to

some attribute or response to an agent or stimulus and these units inherently vary in their response which has nothing to do with measurement error. Here inference about hypothetical parameters is meaningful only in certain special circumstances and even so is merely a limiting case of a predictive inference. In such cases inference (or decision) may be made for a single future observation or several of them jointly or functions of one or more of them depending on the purposes of the investigation. There are special circumstances where the limiting value of the function of the observable, which serves to define a "parameter", may be of interest. When the predictive distribution of a function of M future observables is analytically difficult or too complex to obtain exactly for moderate or large size, the distribution of the limiting value of the function may serve as a convenient approximation for the distribution in the finite case. Sometimes a normative entity is desirable for evaluative and comparative purposes especially when no particular fixed number of observations is necessarily of critical interest. Such a case might rule out all $1 < M < \infty$ and one could restrict one's attention to the case $M = 1$ or $M \to \infty$, the latter of course yields the parametric case. But even in such a situation, if convenient, it is clearly more informative to present the whole spectrum of values for M.

In this predictivistic framework a statistical model indexed by parameters is introduced not because it is necessarily the "true" one. It hopefully serves as an adequate approximation, given what is theoretically assumed and empirically known about whatever the underlying process is that generates the observables. Hence, the paramount issue is not the fictive parameters of a convenient and approximate formulation represented by the parametric model but the potential observables.

The emphasis on observables also has the effect of altering the emphasis in statistics from testing and estimation to model selection and prediction. A predictive analysis may be executed in several different modes which we now elucidate.

The Bayesian Mode

Let the set of random variables $(X_1,\ldots,X_N;X_{N+1},\ldots,X_{N+M})$ or in a more compact notation $(X^{(N)};X_{(M)})$ reflect a partition of past (or to be observed) and future (or to be predicted) variables. Assume that the joint probability function of $(X^{(N)},X_{(M)})$ is

$$f(x^{(N)};x_{(M)}|\ \alpha) = f(x_{(M)}|x^{(N)},\alpha)f(x^{(N)}|\alpha)$$

indexed by a set of parameters α. A prior density for α, $p(\alpha|\beta)$ indexed by β is also part of the structure assumed. For known β the posterior probability function of α, for observed $X^{(N)} = x^{(N)}$, is

$$p(\alpha|x^{(N)},\beta) = \frac{f(x^{(N)}|\alpha)p(\alpha|\beta)}{f(x^{(N)}|\beta)}$$

where

$$f(x^{(N)}|\beta) = \int f(x^{(N)}|\alpha)\ p(\alpha|\beta)d\alpha.$$

The predictive probability function of $X_{(M)}$ is then obtained as

$$f(x_{(M)}|x^{(N)},\beta) = \int f(x_{(M)}|x^{(N)},\alpha)\ p(\alpha|x^{(N)},\beta)d\alpha$$

Hence, any probability statements about the future values $X_{(M)}$ or functions thereof depend on the given probability function and any utilities, losses, costs, etc. that are brought to bear on a specific prediction problem. To illustrate this, consider the case where $(X^{(N)};X_{(M)})$ is a set of independent and identically distributed exponential variables with density function $f(x|\alpha) = \alpha e^{-\alpha x}$ and a prior gamma density for α, $p(\alpha|\delta,\gamma) \propto \theta^{\delta-1}e^{-\gamma\alpha}$. Further, if among the observed $x^{(N)} = (x^{(d)},x^{(N-d)})$, the second set of N-d observations have been censored then it is not difficult to calculate the predictive density of $X_{N+1},\ldots,X_{N+M}$ given $X^{(N)} = (x^{(d)},x^{(N-d)})$

$$f(x_{N+1},\ldots,x_{N+M}|x^{(N)},\ ,\ ) = \frac{\Gamma(x+M+\delta)}{\Gamma(d+\delta)}\ \frac{(s+\gamma)^{d+\delta}}{[s+\gamma+x_{N+1}+\ldots+x_{N+M}]^{d+M+\delta}} \qquad (1)$$

where $s = (x_1 + \ldots + x_N)$. Of course if interest is to be focused on only the next value X_{N+1} we merely set M=1 above to obtain its probability function. We note also that this predictive density is exchangeable, that is, in our assessment the set of future values are exchangeable. In such an instance it is often of interest to calculate the number R of M future X's that lie in some set e.g. $[t,\infty)$, Geisser (1982). If X represented survival time then we might be interested in the fraction that survived until time t. Such a calculation is also easily made since the survival function is

$$\Pr[X \geqslant t \,|\, \alpha] = e^{-\alpha t} = \theta,$$

then

$$\Pr[R = r \,|\, M] = \int \binom{M}{r} \theta^r (1-\theta)^{M-r} \, p(\alpha \,|\, x^{(N)}) \, d\alpha$$

$$= \binom{M}{r} (s+\bar{\gamma})^{d+\delta} \sum_{j=0}^{M-r} \binom{M-r}{j} (-1)^j [s+\bar{\gamma}+t(r+j)]^{-(d+\delta)}$$

Further it can also be shown that as M grows $RM^{-1} \to \theta$ where θ is a random variable whose distribution can be obtained from the distribution of α i.e. $-t^{-1} \log \theta = \alpha$ has the posterior distribution for α whose density is

$$p(\alpha \,|\, x^{(N)}) \propto \alpha^{d+\delta-1} e^{-\alpha(s+\bar{\gamma})}$$

In cases where little is known a priori about α, it is often suggested that $\bar{\gamma} = \delta = 0$ which results in the improper prior density that yields a uniform density for $\log \alpha$. The fiducial approach of Fisher (1956), and the structural approach of Fraser (1968) will yield results which are equivalent to making such an assumption in the Bayesian approach.

Frequency Approach

The classical frequency approach to prediction takes the form of a tolerance region. Here we assume $(X^{(n)};X_{(M)})$ has sampling distribution $F(x^{(N)};x_{(M)}|\alpha)$ with sufficient structure such that $Pr[X_{(M)} \in A(X^{(N)})] = p$ independent of α which represents the chance that the random set $X_{(M)}$ is included in the random region $A(X^{(N)})$. Hence p is the long run relative frequency of the event for random $(X^{(N)};X_{(M)})$, and is interpreted as a measure of the confidence induced in the statement that $X_{(M)}$ will be included in the observed tolerance region $A(x_{(M)})$. These ideas can be applied to the previously discussed exponential example, assuming for simplicity that there is no censoring i.e. $N = d$. Letting $S = \Sigma X_i$ then the sampling distribution of $2\alpha S$ is χ^2_{2N} while $2\alpha X_{N+i}$ is χ^2_2 and all M+1 variables are mutually independent. Transforming to $Z_i = S/X_{N+i}$ $i = 1,\ldots,M$ yields the joint density

$$f(z_1,\ldots,z_M) = \frac{\Gamma(M+N)}{\Gamma(N)} (1 + \sum_{i=1}^{M} z_i)^{-(M+N)}$$

If $M = 1$ then a tolerance interval for the next observation can be obtained through the relationship

$$Pr[X_{N+1} \leq N^{-1} S F_p(2,2N)] = p.$$

For the more general problem the calculation is more difficult. Letting $I = [u,\infty]$, $Z = (Z^{(r)};Z_{(N-r)})$, $z = (z^{(r)};z_{(N-r)})$, $\int_{I(z^{(r)})}$ and $\int_{I^c(z_{(N-r)})}$ the r and N-r fold integrals where $z_1,\ldots,z_r$ and $z_{r+1},\ldots,z_n$ are integrated over I and I^c respectively, then

$$P_r = Pr[\text{exactly } R=r \text{ of } M \text{ Z's} \in I) = \binom{M}{r} \int_{I(z^{(r)})} \int_{I^c(z_{(N-r)})} f(z^{(r)};z_{(N-r)}) dz(r) dz_{(N-r)}$$

The right hand side is a function of r, M and N. For a tolerance interval on R or say $R \leq r_0$ one computes $\sum_{r=0}^{r_0} P_r = Q_a$ at the value $a = tN^{-1}s$ where s is the observed value of $S = X_1 + \ldots + X_N$. Because of the form of the previous

calculation it is clear that tolerance intervals for this case may be calculated from the previous Bayesian calculation by virtue of setting $\gamma = \delta = 0$ and transforming $X_{N+i} = Z_i s$, $i = 1,\ldots,M$ in (1).

In the frequency mode a highly distribution-robust procedure is also available. We need only assume that the underlying distribution of a set of exchangeable variables $X_1,\ldots,X_N$ is absolutely continuous (note that independence is not even necessary).

Let the ordered values of $X_1,\ldots,X_N$ be $X_1' \leq X_2' \leq \ldots \leq X_N'$, then for the interval

$$I_{jk} = (X_j', X_{j+k}'),$$

defining $I_{j,N+1-j} = (X_j', \infty)$ and $I_{0,k} = (-\infty, X_k')$
it can be shown by combinatorial methods, Wilks (1962), that

$$\Pr[\text{exactly } R=r \text{ out of } M \; X_{N+i}\text{'s lie in } I_{jk}] = \frac{\binom{k+r-1}{r}\binom{N+M-k-r}{M-r}}{\binom{N+M}{M}} = P_{r,k}$$

and

$$\Pr[\text{exactly } R=r \text{ out of } M \; X_{N+i}\text{'s exceed } X_j'] = \frac{\binom{N+r-j}{r}\binom{M-r+j-1}{M-r}}{\binom{N+M}{M}} = P_{r,N+1-j}$$

so that

$$\Pr\left[\frac{R}{M} \leq \frac{r}{M}\right] = \sum_{x=0}^{r} P_{x,N+1-j}.$$

In this case, only probabilities for fractions exceeding order statistics can be exactly computed. For the special case of a single future observation X_{N+1}, the result is simply

$$\Pr[X_j' < X_{N+1} < X_{j+h}'] = \frac{k}{N+1}. \qquad (2)$$

$$\text{(5)} \qquad \cdots$$

Other modalities

There are several modes that have been suggested for scaling future values which depend on sufficiency and relative likelihood concepts. The relative likelihood procedure, Fisher(1956), for scaling future values depends on $X^{(N)}$ and $X_{(M)}$ being independent given α and defining

$$L(\hat{\alpha}_N|x^{(N)}) = \sup_{\alpha} L(\alpha|x^{(N)}), \quad \hat{\alpha}_N = \hat{\alpha}(x^{(N)});$$

$$L(\hat{\alpha}_M|x_{(M)}) = \sup_{\alpha} L(\alpha|x_{(M)}), \quad \hat{\alpha}_M = \hat{\alpha}(x_{(M)});$$

$$L(\hat{\alpha}_{N+M}|x^{(N)},x_{(M)}) = \sup_{\alpha} L(\alpha|x^{(N)},x_{(M)}), \ \hat{\alpha}_{N+M} = \hat{\alpha}(x^{(N)},x_{(M)}).$$

Then the function

$$RL(x^{(N)}|x_{(M)} = \frac{L(\hat{\alpha}_{N+M}|x^{(N)},x_{(M)})}{L(\hat{\alpha}_N|x^{(N)})\ L(\hat{\alpha}_M|x_{(M)})}$$

is used to scale the plausible values of $x_{(M)}$ for observed $x^{(N)}$. A similar function is obtained when $S_N = S(X^{(N)})$ and $S_{N+M} = S(X^{(N)}, X_{(M)})$ are sufficient for θ based on the observed set and the total set respectively. By the properties of sufficiency one obtains the conditional probability function of S_N given $S_{N+M} = s_{N+M}$,

$$f(s_N|s_{N+M}) = prlk(x_{(M)}|x^{(N)}),$$

to be independent of α. The above is then used to scale values of $x_{(M)}$ given $x^{(N)}$, Lauritzen (1974). For the simple exponential example previously given we obtain the following scaling functions for x_{N+1},

$$RL(x^{(N)}|x_{N+1}) \propto \frac{x_{N+1}}{(N\bar{x} + x_{N+1})^{N+1}},$$

and

$$prlk(x^{(N)}|x_{N+1}) \propto \frac{1}{(N\bar{x} + x_{N+1})^{N}}.$$

A predictive mode which makes no distributional assumptions termed predictive sample reuse PSR, Geisser (1974, 1975) requires the following ingredients:

1. An arbitrarily chosen predictive function of a future observable

$$x = x(x^{(N)}, \alpha) \qquad \alpha \in A$$

where α is a set of values to be determined;

2. A schema $S = S(N,n,\Gamma)$ of partitions where

$$P_i = (x_{ir}^{(N-n)}, x_{io}^{(n)})$$

is the i^{th} partition of $x^{(N)}$ into $x_{ir}^{(N-n)}$ the set of N-n retained values and $x_{io}^{(n)}$ the set of n omitted values of $x^{(N)}$. The defined set of such partitions for a given n is Γ, say, and the number of such partitions $P_i \in \Gamma$ is P. The predictive function is applied to the retained observation set $x_{ir}^{(N-n)}$ and used to predict $x_{io}^{(n)}$, the deleted set for each P_i yielding $\hat{x}_{io}^{(n)}(\alpha)$ which is a function of α.

3. A discrepancy measure

$$D_n(\alpha) \propto \sum_{P_i \in \Gamma} d(x_{io}^{(n)}, \hat{x}_{io}^{(n)}(\alpha))$$

where d(a,b) is a defined measure of the discrepancy between two n-dimensional vectors.

4. To obtain the final predictor $\hat{x} = x(x^{(N)}, \hat{\alpha})$ of the future value, $D_n(\alpha)$ is minimized with respect to α which yields $\hat{\alpha}$.

As a simple example consider as a predictive function a linear combination $x = \alpha h + (1-\alpha)m$, $\alpha \in [0,1]$, of the median m and the average of symmetric order statistics

$$h = \tfrac{1}{2}(x'_{[pN+1]} + x'_{N-[pN]}),$$

where [m] represents the largest integer in m and $0 < p < .5$, for N observations. Use of a squared error discrepancy based on a one-at-a-time omission schema requires minimization of

$$D_1(\alpha) \propto \sum_{j=1}^{N} (\alpha h_j + (1-\alpha)m_j - x_j)^2$$

where h_j and m_j are h and m respectively with x_j deleted. The solution

yields for the predictor

$$x = h \qquad \text{if } \hat{\alpha} \geq 1$$

$$= \hat{\alpha}h + (1-\hat{\alpha})m \qquad \text{if } 0 < \hat{\alpha} < 1$$

$$= m \qquad \text{if } \hat{\alpha} \leq 0$$

where

$$\hat{\alpha} = \sum_{j=1}^{N}(h_j - m_j)(x_j - m_j) / \sum_{j=1}^{N}(h_j - m_j)^2.$$

Sample reuse intervals may also be obtained using similar ingredients, Butler and Rothman (1980). They are

1. A predictive interval function

$$P.I.(x^{(N)};\alpha)$$

2. A criterion function (assuming a simple one-at-a-time omission schema)

$$D_1(\alpha) \propto \sum_{j=1}^{N} L\{P.I.(x_j^{(N-1)};\alpha)\} \qquad (3)$$

where $L\{\cdot\}$ is defined as the length of the j^{th} interval based on all of the observations but the j^{th}.

3. A relative frequency of coverage of $1-\beta$ in a predictive simulation is then obtained by minimizing $D_1(\alpha)$ with respect to α subject to

$$\frac{1}{N}\sum_{j=1}^{N} I[x_j \notin P.I.(x_j^{(N-1)};\alpha)] \leq \beta, \qquad (4)$$

where I is the indicator of the event in brackets.

The resulting solution for $\hat{\alpha}$ is then substituted in the predictive interval function to obtain

$$P.I.(x^{(N)};\hat{\alpha}).$$

As a very simple illustration, consider the predictive interval function which uses the symmetric order statistics

$$P.I.(x^{(N)};\alpha) = (x'_{\alpha}, x'_{N+\alpha+1}).$$

Minimizing the criteria function (3) subject to the constraint (4) and setting $\beta = 2p$ we obtain as solution

$$P.I.(x^{(N)};\hat{\alpha}) = (x'_{[Np]}, x'_{N+1-[Np]})$$

with coverage 1-2p. For $p = \frac{1}{N}$ the simulated coverage is $\frac{N-2}{N}$ that the $N+1^{st}$ observation lies within the range of the previous N observations. If we compare this with the result (2) of the more structured situation for which the tolerance coefficient is $\frac{N-1}{N+1}$ it is clear that it is as if the loosening of the structure manifests itself in the loss of a single observation.

Customary Applications

In sample surveys, the problem is to estimate some function of a finite number of observables - part observed and part unobserved. Clearly then this is a prediction problem, even if the function is sometimes misdesignated as a parameter. Direct prediction problems as such, abound in Regression, Time series, Growth Curves, Lee and Geisser (1972), and a variety of other special topics where the modeling clearly anticipates the need for prediction. A few less direct areas will be discussed in some detail.

Probability "Estimation"

One immediate application is to the so-called density estimation problem or the estimation of the distribution function. Clearly in the Bayesian mode, the predictive distribution (density), which is the expectation of the sampling distribution (density) over the posterior distribution of the parameters, is, for squared error loss, the optimal estimator of the sampling distribution (density). Other loss functions will lead to other estimates, c.f. Geisser (1982). Hence probability estimation and whatever is derived from it is contained in this approach - so that even such problems as "goodness of fit" can be managed in this way, Guttman (1966), Geisser (1971).

Classification

Classification problems are essentially prediction or, in point of time, retrodiction problems. For the sake of simplicity consider two populations π_1 and π_2 with training samples $D_1 = (x_1,\ldots,x_n)$ and $D_2 = (y_1,\ldots,y_m)$ and a new observation $Z = z$ which has known prior probability p_i of originating from π_i, $i = 1,2$ and $p_1 + p_2 = 1$. Assume π_i is specified by densities $f_i(\cdot|\theta_i)$ and $p(\theta_1,\theta_2)$ is the assumed prior probability function for the set of parameters (θ_1,θ_2). For $D = (D_1,D_2)$, the posterior probability for the origin of z is

$$\Pr[\pi_i|z] \propto p_i f(z|D,\pi_i)$$

where

$$f(z|D,\pi_i) = \int f_i(z|\theta_i)\, p(\theta_1,\theta_2|D)\, d\theta_1 d\theta_2$$

Classification of Z may be made to that population which maximizes the posterior probability if there is no differential cost. For multivariate normal applications see Geisser (1964, 1966).

Model Selection

Let M_i be a model, $i = 1,\ldots,k$ which specifies the probability function for a set of observations D to be $f(D|M_i,\theta_i)$, indexed by unknown θ_i. Then, for p_i the prior probability of model M_i the posterior probability of M_i is

$$\Pr[M_i|D] \propto p_i f(D|M_i)$$

where

$$f(D|M_i) = \int f(D|M_i,\theta_i)\, p(\theta_1,\ldots,\theta_k)\, d\theta_1 \ldots d\theta_k$$

is the predictive (marginal) density of the observation set D and $p_1 + \ldots + p_k = 1$. Again, in the absence of other considerations, the maximum $\Pr[M_i|D]$ could be used to select the most appropriate model. For variations on this theme see Geisser and Eddy (1979).

Problems of Comparison

The comparison of certain attributes of groups or populations comprises a major portion of the statistical enterprise. Current practice often dictates that certain location parameters be made the focus of comparison. For example, in a modeling which posits two normal populations $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$, the focus may be on inferential statements about $\eta = \mu_1 - \mu_2$, based on samples $D_i = (x_{i1}, \ldots, x_{iN_i})$ $i = 1,2$. For example, a Bayesian would base his inference on the posterior distribution $P(\eta | D_1, D_2)$.

A predictive comparison which includes this as a limiting case would focus on $Z = \bar{Z}_1 - \bar{Z}_2$ where $\bar{Z}_i = M_i^{-1}[Z_{i1} + \ldots + Z_{iM_i}]$, $i = 1,2$ and display the predictive distribution and density $F(z | D_1, D_2)$ and $f(z | D_1, D_n)$ respectively. Notice for $M_1 = M_2 = 1$, we are comparing the distribution of the difference of two observations one drawn from each population. As M_i grows $Z \to \eta$, so the former parametric analysis is the limiting case of the latter, but it is quite likely that interest would be focused on a finite number of future values except for a normative evaluation. At any rate the predictive comparison is richer and more informative. A variety of comparison problems can be handled from the predictive point of view in particular optimal ranking and selection problems, Geisser (1971).

Regulation and Optimization

In problems of regulation - where a series of N trials or experiments are made indexed by $t \in T$ resulting in (t_i, x_i), $i = 1, \ldots, N$, - the object is to produce a value in a set X_0 by appropriate choice of t. Closely allied to this is the optimization problem which requires selecting t to yield an optimal but unknown future value for x (for example a minimum or maximum). If the future experiment was already performed and x observed but the index t was unknown and required identification then a calibration problem results.

In all of these cases the key to the solution within a Bayesian framework is the predictive distribution of a future X, c.f. Aitchison and Dunsmore (1975).

Model Criticism

In cases where alternative models are not available a predictive analysis may also be useful in criticizing an entertained model. Suppose within a Bayesian framework the model consists of observable X and parameter set θ structured as follows

$$p(x,\theta) = p(x|\theta)\, p(\theta).$$

By computing the marginal predictive probability function

$$p(x) = \int p(x,\theta)\, d\theta$$

there exists the potential to assess the credibility of the entertained model for an observed set $X = x$. A simple predictive signifiance test can be defined at level α by setting

$$\alpha = \Pr\{p(X) < p(x)\}$$

thus allowing criticism of at least some aspects of the model, c.f. Aitchison and Dunsmore (1975), Box (1980). Although this procedure allows questioning the model as a whole - it may reject a model merely because of one or a few spurious observations. These potential offenders may be pinpointed by calculating conditional predictive diagnostics such as

$$p(x_j | x_{(j)})$$

where $x_{(j)}$ represents all of the observations except for x_j. Those x_j which yield relatively small values could indicate precisely where the difficulty lies, Geisser (1980).

Influential Observations

Other methods particularly useful for regression analysis in characterizing and detecting the influence of observations singly or in sets on prediction have been developed by Johnson and Geisser (1982, 1983). This method compares the predictive probability functions of future observations, f and $f_{(i)}$, with and without the observation(s) respectively, to determine the effect or influence of that observation(s) on prediction. Although others may be used, in particular a scalar measure of the effect that is found useful, is the Kullback-Leibler divergence measure

$$I_{(i)} = E_{f_{(i)}} \log(f_{(i)}|f).$$

Each observation (or subset of fixed size) is then ranked according to $I_{(i)}$ to determine its relative effect on the predictive distribution. Once influential observations have been identified, it is up to the practitioner to decide what action, if any, to take with respect to them. For a detailed analysis of such a situation see Johnson and Geisser (1983).

In summary, almost all areas of statistical application can be informatively managed by a predictivistic approach.

References

Aitchison, J. and I.R. Dunsmore (1975) Statistical Prediction Analysis Cambridge University Press, New York, Melbourne.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society A, 143, 383-430.

Butler, R. and E.D. Rothman (1980). Predictive intervals based on reuse of the sample. Journal of the American Statistical Association, 75, 372, 881-889.

Fisher, R.A. (1956). Statistical Methods and Scientific Inference, 1st edition. Hafner, New York.

Fraser, D.A.S. (1968). The Structure of Inference John Wiley, New York.

Geisser, S. (1964). Posterior odds for multivariate normal classification, Journal of the Royal Statistical Society B, 1, pp. 69-76.

Geisser, S. (1966). Predictive discrimination, Multivariate Analysis, edited by P. Krishnaiah, Academic Press, New York, pp. 249-163.

Geisser, S. (1971). The inferential use of predictive distribution, B.P. Godambe and D.A. Sprott, eds., Foundations of Statistical Inference Holt, Reinhardt and Winston, Toronto, Montreal pp. 456-469.

Geisser, S. (1974). A predictive approach to the random effect model, Biometrika 61, 101-107.

Geisser, S. (1975). The predictive sample reuse method with applications, Journal of the American Statistical Association 70, 350, 320-328.

Geisser, S. (1980). Discussion of Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, A, 143, 416-417.

Geisser, S. (1982). Aspects of the predictive and estimative approaches in the determination of probabilities, Biometrics (supplement), 38,1 75-93.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. Journal of the American Statistical Association, 74, 153-160.

Guttman, I. (1966). The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society, B, 29, 83-100.

Johnson, W., and Geisser, S. (1982). Assessing the predictive influence of observations, Essays in Honor of C.R. Rao, Ed. Kallianpur, Krishnaiah, and Ghosh, Amsterdam: North Holland, 343-358.

Johnson, W., and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis, Journal of the American Statistical Association, 78, 381, xxx-yyy.

Lee, J.C. and Geisser, S. (1972). Growth Curve Prediction. Sankhyā A, 34, 393-412.

Lauritzen, S.L. (1974). Sufficiency, prediction and extreme models, Scandinavian Journal of Statistics, 1, 128-134.

Pearson, K. (1907). On the influence of past experience on future expectation. Philosophical magazine, XIII, sec. 6, 365-378.

Wilks, S.S. (1962). Mathematical Statistics, John Wiley and Sons, New York.